NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY

Beginning a Higher Trajectory: Grade 11 Study

State Teachers of the Year Compare Former and New State Assessments

National Network of State Teachers of the Year

Catherine McClellan, Ph.D.
Jilliam Joe, Ph.D.
Katherine Bassett, M.Ed.

March 2017

NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY

The National Network of State Teachers of the Year is pleased to share with you our latest in a series of reports on high-quality summative student assessment. In our preceding two reports, *The Right Trajectory* and *Still the Right Trajectory*, we examined the grade 5 assessments previously used by individual states and compared them with the two consortia assessments.

In this report, we continue our focus on the important issue of assessing our students' learning through standardized, summative assessments. Using research-based methodologies and practices, and survey instruments designed for this study, we convened a panel of twelve outstanding educators to examine the Smarter Balanced (SBAC) grade 11 assessment. The study panel included State and National Teachers of the Year and Finalists for State Teacher of the Year.

The study methodology for this case is significantly different from the methodology used in our previous studies. In this case, the panel examined only the Smarter Balanced grade 11 assessment because there were no previous state assessments that could be used for comparisons. As we did in our previous two reports, we used the Smarter Balanced assessment supplied to us by the consortium for this study. Smarter Balanced is a computer-adaptive test, but it was important for the study that teachers examine the same form, so we used a form fixed at the 60th percentile of student performance.

Working with our study partners, EducationCounsel (on the policy side) and Clowder Consulting (on the science end), we are eager to share our results. In this case, the teachers found that:

- The new consortium assessment reflects an appropriate depth and range of content.
- The distribution of the consortium assessment's content, while representative, does not fully encompass excellent 11th grade instruction.
- The new consortium assessment measures concepts learned in the classroom and promotes curriculum-centered test preparation.
- Though an improvement, the new 11th grade consortium assessment is not yet sufficiently rigorous nor cognitively complex.

These educators found many things to like about the assessment, and they made recommendations for continuous improvement. We are eager to share their counsel with you.

An NNSTOY core belief is that educators should always be at the table when education policy is being crafted, debated or modified. We are excited to share our findings with you and look forward to working with you in bringing the voice of educators to the policy process.

With warm regards,

Katherine Bassett

Acknowledgements

NNSTOY wishes to thank the following individuals and groups for their support and contributions to this project:

Our Partners

Our partner in this work, EducationCounsel, specifically Mr. Scott Palmer, Mr. Nick Spiva and Ms. Sandi Jacobs, for their commitment to learning what a group of outstanding educators thought about the trajectory that we are taking in moving to new state assessments. Their guidance, policy expertise and assistance with access to the assessments studied was invaluable, as was their overall collaboration.

Our science partner in this work, Clowder Consulting. Dr. Catherine McClellan is a consummate psychometrician and research scientist. Her vast knowledge of survey science, research methodology and analytic ability made this research study possible. Dr. Jilliam Joe is a gifted facilitator of focus groups, and her analytic capabilities made unpacking data understandable and clear for lay people.

We thank both sets of partners for their patience, dedication and collaboration in this lengthy process.

Assessment Providers

Allowing an outside agency access to confidential assessment material is a serious undertaking. We are most grateful to Smarter Balanced for giving us access to their assessment. We protected the confidentiality of this assessment diligently and appreciate your allowing us access to them. Without this access, there would be no study.

Our Funders

We were fortunate to have generous funding with which to conduct this study supplied by the Rockefeller Philanthropy Advisors and the Bill and Melinda Gates Foundation. Without this funding, this study would not have taken place. We are most grateful.

Our Reviewers

We sincerely thank the following for making the time to conduct an external review of this report: Mr. Chris Minnich, Executive Director of the Council of Chief State School Officers; Dr. Rebecca Snyder, Pennsylvania State Teacher of the Year 2009 and Past President, NNSTOY; Dr. Joshua Starr, Chief Executive Officer, PDK.

The Panelists

Finally, we could not have asked for a more prepared and committed set of educators with whom to do this work. Each panelist made certain to be well-prepared for the work of the study. Each is an exemplary educator and brought intense knowledge, skill and ability to the table. Each entered into this work without preconceived ideas or opinions about the assessments. Each is a shining example of the best in education in our country and we are grateful for their participation.

Table of Contents

Beginning a Higher Trajectory: Grade 11 Study

This study continues the work that NNSTOY and its partners, Clowder Consulting and Education-Counsel, began with our *Right Trajectory* study, released in 2015, and continued in our *Still the Right Trajectory* study released in early 2017. In addition to these studies examining the prior 5th grade state tests and the consortium assessments, this research includes analysis of the 11th grade Smarter Balanced assessment. Although our research is based on the same survey used for the 5th grade study, the design for the 11th grade research was fundamentally different because reviewers only examined the Smarter Balanced test, without using a prior test as a basis for comparison. Twelve expert teachers from Department of Defense, Idaho, Minnesota, Mississippi, Nebraska, Nevada, Utah and Washington participated in the panel, and each panelist evaluated either the reading/English language arts (ELA) or mathematics portion of the assessment. Despite the limitations imposed by the lack of context from prior tests, the impressions of outstanding educators about the strengths and areas for improvement of the 11th grade assessment provide an important contribution to the ongoing conversation about the value of standardized assessments. This analysis presents both positive and constructive feedback for consideration as test developers and states reflect and engage in continuous improvement.

Executive Summary

1. **The new consortium assessment reflects an appropriate depth and range of content.** Teachers in our study spent time carefully examining the Smarter Balanced assessment at the 60th percentile for 11th grade students. A majority of the teachers (67%) formed the impression that the assessment measured knowledge, skills and abilities that are appropriate based on the framework used for 11th grade students in both range and depth. In addition, they found that the content was neither above nor below grade level. (A smaller percentage, 33%, suggested there were not enough items that were above grade level.)

1 Under No Child Left Behind, states had only to assess students one time in high school. Most states, including those with panelists in this study, did not have assessments specific to grade 11 to use as a basis of comparison. No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

2 The assessment examined is a test form that a student would see when his score places him above 60% of the students in the population taking the test.

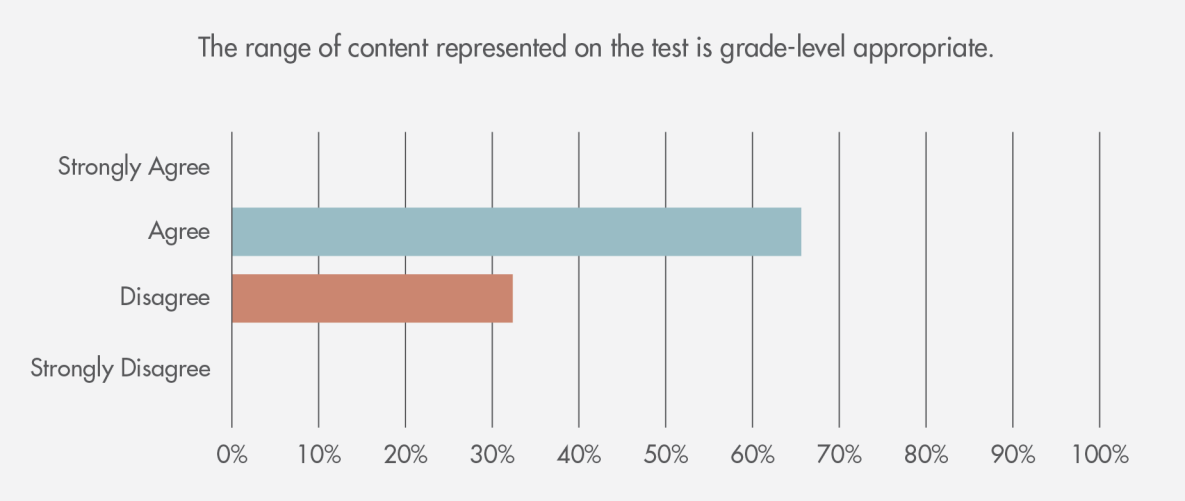


Figure 1. Percent agreement with statement: "The range of content represented on the test is grade-level appropriate."

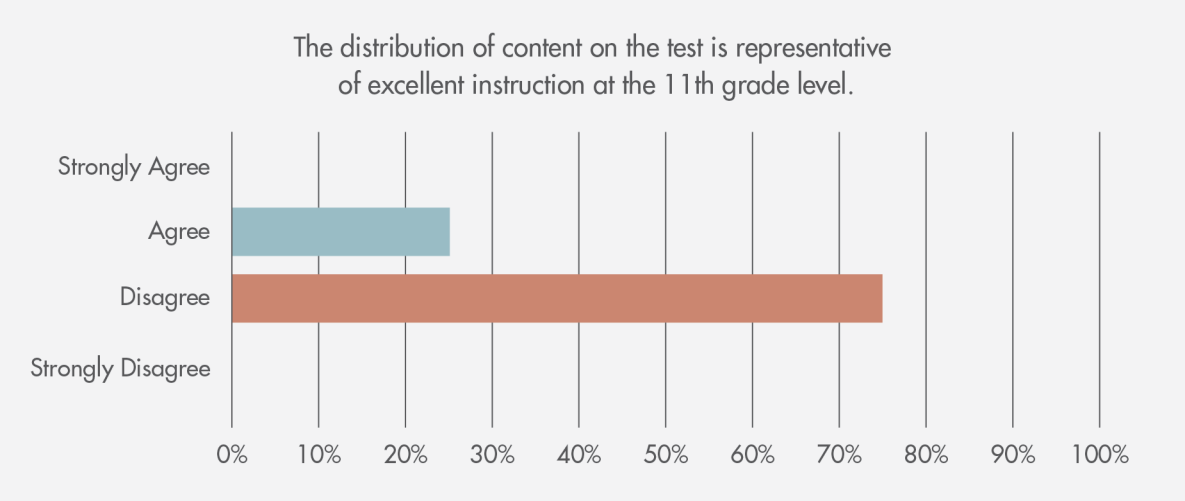


Figure 2. Percent agreement with statement: "The distribution of content on the test is representative of excellent instruction at the 11th grade level."

2. **However, the distribution of the consortium assessment's content, while representative, does not fully encompass excellent 11th grade instruction.** Teachers perceived that the assessment did not cover a wide enough range of content and cognitive challenge to fully reflect excellent instruction at the 11th grade level. In the case of ELA, for example, teachers noted that the test seemed to favor one genre of literature over others. One teacher said, "The students are doing far more nonfiction on the test than fiction, [and fiction] is predominantly what happens in English classrooms." Others noted: "I don't feel like this sampling of questions reflects all of the essential topics in an 11th grade curriculum. I would have liked to see additional trigonometry and additional parent families of functions represented" and "I ask a lot more of my students than this test demonstrates." While the Common Core State Standards emphasize non-fiction and specific mathematics content at grade 11, the teachers found that the assessment was not entirely consistent with the mix of source materials and subject matter taught in excellent 11th grade classrooms.

3. **The new consortium assessment measures concepts learned in the classroom and promotes curriculum-centered test preparation.** As a matter of validity, it is important that an assessment measures what it purports to measure. Variability in test scores should

largely be attributed to true differences in content knowledge and mastery rather than some outside factor, like test-taking ability. A strong majority (83%) of the teachers agreed that the 11th grade assessment addressed this, as seen below. Another encouraging finding from this study is the extent to which teachers perceived that instruction going beyond skill-and-drill strategies was necessary to prepare students to be successful on the assessment: 67% believed that it was.

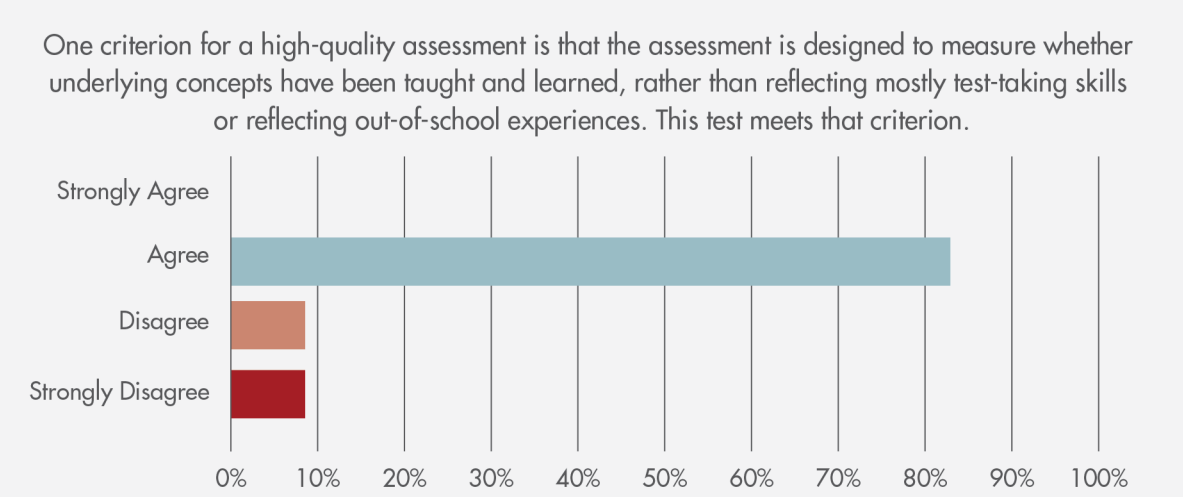


Figure 3. Percent agreement with statement: "One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion."

4. **Though it is an improvement, the new 11th grade consortium assessment is not yet sufficiently rigorous nor cognitively complex.** The increased investment in rigorous and cognitively demanding instruction designed to prepare students for 21st century college and career success was not fully mirrored on this assessment. A majority of teachers did not think the test requires students to demonstrate important high-level thinking skills like experimentation, analysis and synthesis. One teacher stated, "I think, as an 11th grade teacher, we talk more about the process and not just the end product." In this teacher's estimation, the test did not go far enough to measure the thinking that generates a student's response.

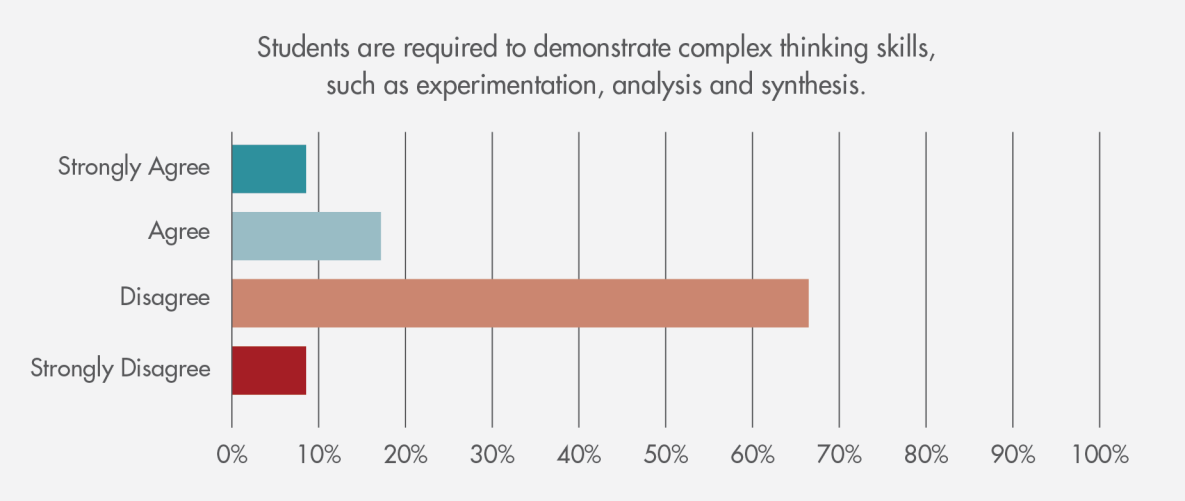


Figure 4. Percent agreement with statement: "Students are required to demonstrate complex thinking skills, such as experimentation, analysis and synthesis."

Summary

The parameters of the study did not allow for comparison of the new assessment to other state assessments. Examination of the 11th grade Smarter Balanced assessment produced mixed opinions. Panelists believed that the assessment was appropriate for 11th grade students and required that students understand the content in order to perform well. It is difficult to develop an assessment that is challenging for all students without being overwhelming. However, the teachers indicated that the assessment did not fully sample the instruction in an excellent 11th grade classroom and that it should require more complex thinking.

Overview of Study

The 11th grade Smarter Balanced assessment was selected for this study. The assessment represents an evaluation of high school content and skills and is part of the testing required for Adequate Yearly Progress (AYP) metrics. Five key questions were investigated:

1. **Does the new consortium assessment reflect the range of knowledge and skills that all students should know?**
2. **Is the new consortium assessment designed to reflect the full range of cognitive complexity in a balanced way?**
3. **Does the new consortium assessment align with the strong instructional practices these teachers use in the classroom, and thereby support great teaching and learning throughout the school year?**
4. **Does the new consortium assessment provide information relevant to a wide range of performers?**
5. **While the new consortium assessment is more rigorous and demanding, is it grade-level appropriate?**

Methodology[3]

The study was organized in two phases.

- The first phase comprised an in-depth review using Webb's Depth of Knowledge (DOK) framework. In preparation for the alignment work to come, the panelists participated in an online webinar exposing them to DOK. In addition, each panelist was asked to prepare for the study panels by reviewing their own state's standards in math and English Language Arts (ELA).
- The second phase comprised an evaluation of the new assessment and whole-group discussion of selected results. A panel of teachers examined the new Smarter Balanced 11th grade consortium assessment. The review took place over a day and a half. Two instruments administered online were used along with a collection of demographic data. All instruments underwent several reviews prior to their final use. (The surveys are provided in Appendix B.)

  In order to gather more detailed information about aspects of the assessment and the reasons the panelists had for their responses to survey items, a focus group discussion of the assessment was held after the reviews and surveys were completed.

  The participants were given an orientation to the assessments they would review. During the orientation, they were encouraged to work through the items as if each were a typical well-prepared 11th grade student, not necessarily the kind of student who happened to be in their individual classrooms. This calibration provided a common lens through which to evaluate the cognitive demand associated with a particular assessment item.

  There were some contextual and design differences between this and the previous 5th grade studies with which readers may be familiar. First, and likely most important, the 11th grade panel had no other 11th grade state tests to compare to the Smarter Balanced assessment. It may be that this lack of context and the sense of progress (or lack thereof) had an impact on the evaluation of the 11th grade assessment. Evaluating a single test in isolation is qualitatively different than comparing or considering three tests as a set.

  Additionally, the 11th grade review was hampered by some technical difficulties. At the beginning of the review, the online system was not displaying the graphics and media associated with the items properly. This issue was resolved after a couple of hours, but the problem may have affected the panelists' judgment of the quality of the items and/or the assessment.

3 For a more-detailed description of the procedures used in this study, see "The Right Trajectory" report pages 14 and 15, available at http://www.nnstoy.org/wp-content/uploads/2015/11/Right-Trajectory-FINAL.pdf

K-12 Assessments and Survey Instruments

Smarter Balanced

The Smarter Balanced consortium assessment was designed to measure the standards set forth by the CCSS. It is typically administered to students as a computer-adaptive test (CAT). However, we did not use the CAT version of the test for the purposes of the study. The assessment evaluated was a linear form based on a student at the 60th percentile of the proficiency distribution at grade 11. There were 43 selected-response, short- and extended-response items on the ELA assessment that comprised reading and listening passages. The math assessment comprised 48 items, including selected-response, short- and extended-response items as well as one extended stimulus passage on which several items were based.

Survey Instruments

Two online survey instruments were developed for this study. *The Attitudes Toward Tests* survey was designed by the research team to capture teachers' perceptions about tests and item types. Educators may hold preferences for how best to measure student knowledge and skills. We think it was important to understand what these preferences were for participants prior to and after engaging with the assessments.

The *Survey of Assessment Quality* was developed to evaluate the five key areas of quality of the assessments listed above. These items address the appropriateness and rigor of the items for low-, mid- and high-performing students; the content; performance levels; balance; and grade-appropriateness of the items in each of the assessments overall. In addition, a background questionnaire was created to gather relevant demographic and background information such as school type, years of experience and content areas, about participants. All instruments underwent several reviews prior to their final use.

Participants

We convened 12 outstanding educators for the study, all State Teachers of the Year and Finalists recognized for excellence in classroom practice. The panel was, to the extent feasible, diverse on the following dimensions:

- **Content area**. We selected panelists with rich teaching experience in either Math or ELA.
- **States.** Participants included teachers from Idaho, Minnesota, Mississippi, Nebraska, Nevada, Utah and Washington. There was also one teacher from the Department of Defense.
- **Race/ethnicity and gender.** We sought to reflect the racial/ethnic and gender makeup of the general teaching population to the extent possible.
- **School setting.** We worked to bring together panelists from a variety of school settings, e.g. rural, suburban, urban.

We selected teachers who are familiar with 11th grade instruction. (More detailed demographic data on the panelists is presented in Appendix A.) For taking part in this study, participants were given a stipend for their time and reimbursed for expenses incurred for travel, lodging and food. No other compensation was provided.

---

1 OAKS and Smarter Balanced are adaptive tests, but teachers only reviewed one linear form based on a student at the 60th percentile of the proficiency distribution at 5th grade.

Data Collection

The review process drew on participating teachers' existing areas of expertise to determine how well the assessments reflect the kind of teaching and learning that they want to see in the classroom. The panel met in Las Vegas, Nevada, for two days of onsite activities in early August. We employed a four-step data gathering process. These steps were:

1. Training and orientation (including the *Attitudes Toward Tests* survey)
2. Webb DOK alignment
3. Assessment review using the *Survey of Assessment Quality*
4. Focus group discussion

Results

This section provides insight into the results of the findings from the *Survey of Assessment Quality*. Herein the results will be organized under three groupings: positive reactions, mixed or neutral opinions, and negative or critical reactions. The complete set of response data can be found in Appendix B. Qualitative data from the follow-up discussions are integrated with our summary of the survey data, where appropriate, for clarification and illumination. (Discussion prompts are provided in Appendix C.)

Positive Reactions

Panelists were asked the extent to which they agreed with the statements: "The depth of content represented on the test is grade-level appropriate" and "The range of content represented on the test is grade-level appropriate." The results are shown in Figures 5 and 6. The majority of panelists agreed that both the depth and range of content on the Smarter Balanced tests were appropriate for grade 11. These items offer support for a positive answer to key questions 1 and 5.

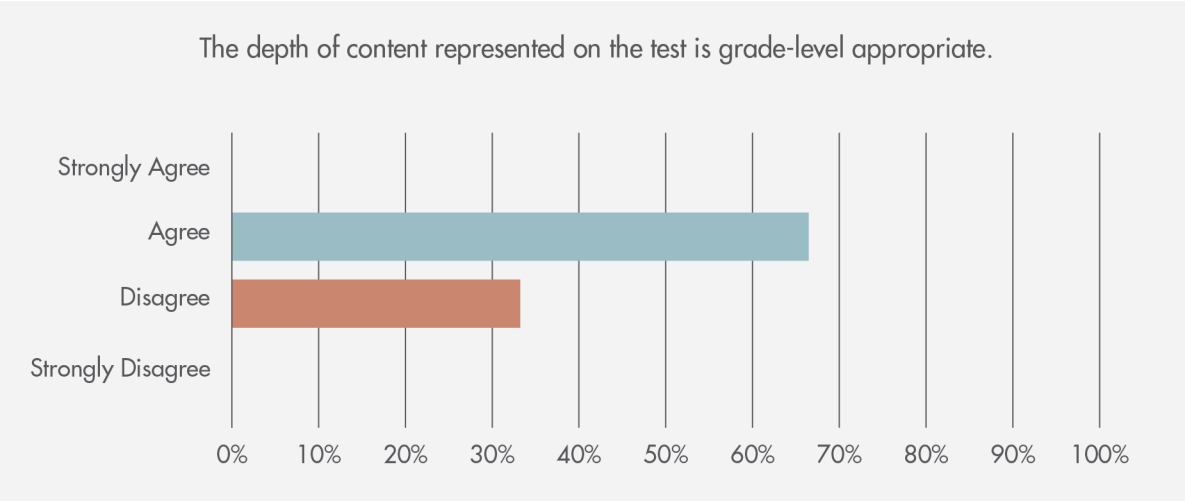


Figure 5. Percent agreement with statement: "The depth of content represented on the test is grade-level appropriate."

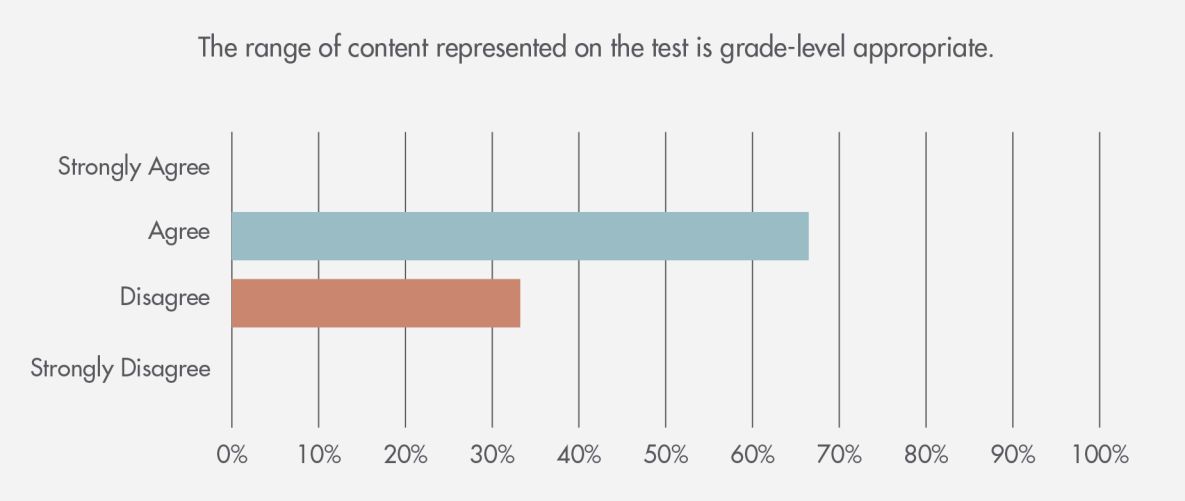


Figure 6. Percent agreement with statement: "The range of content represented on the test is grade-level appropriate."

Teachers on the panel were asked to rate their agreement with the statement: "One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or out-of-school experiences. This test meets that criterion." The results are shown in Figure 7, with a large majority of teachers agreeing with the statement.

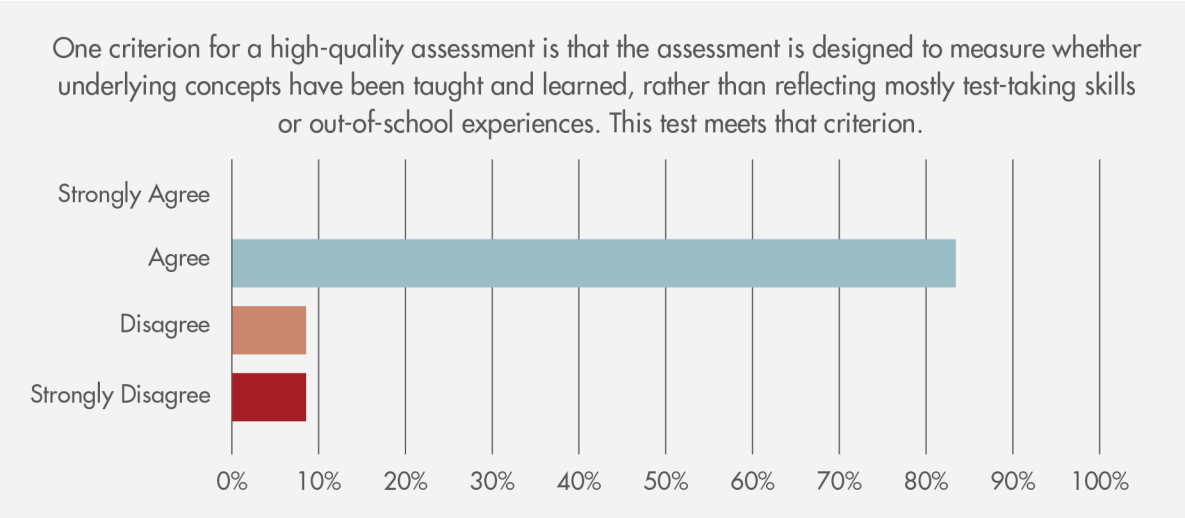


Figure 7. Percent agreement with statement: "One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion."

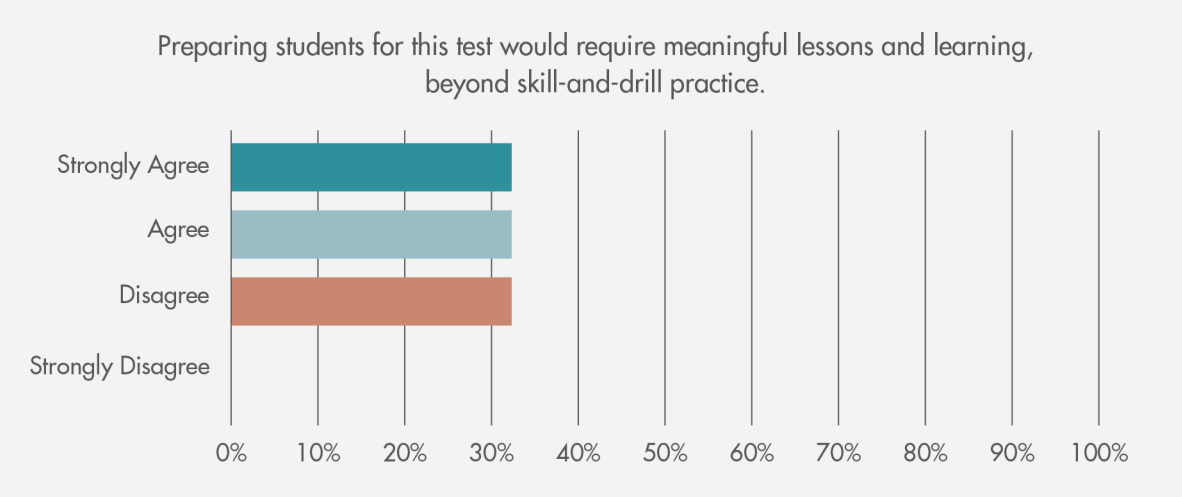


Figure 8. Percent agreement with statement: "Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice."

When asked about their agreement with the statement: "Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice," there was less total agreement, though more teachers said they Strongly Agree with the statement.

The responses to two items on formative assessment were generally positive, with 75% combined agreement, as can be seen in Figures 9 and 10.

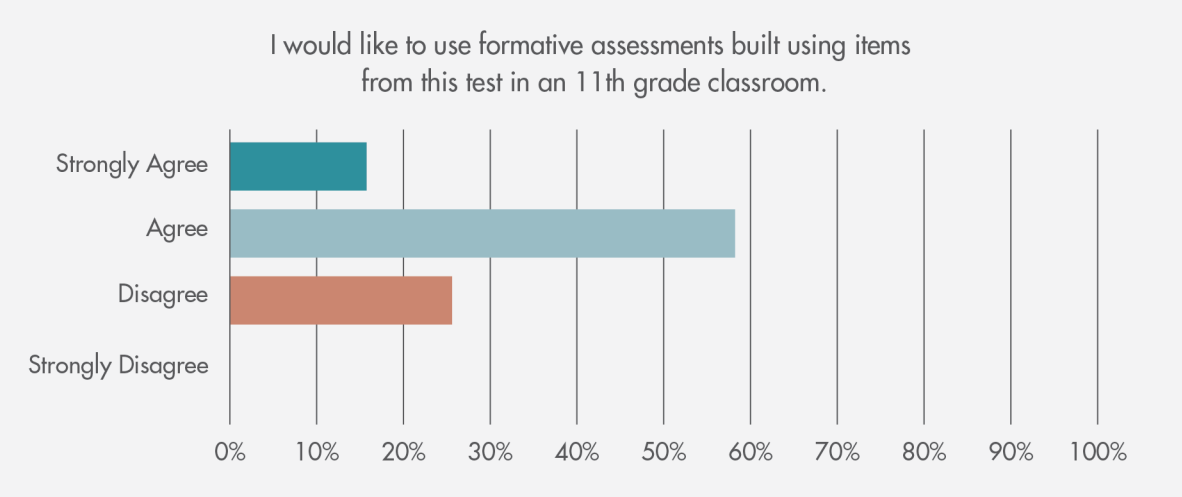


Figure 9. Percent agreement with statement, "I would like to use formative assessments built using items from this test in an 11th grade classroom."

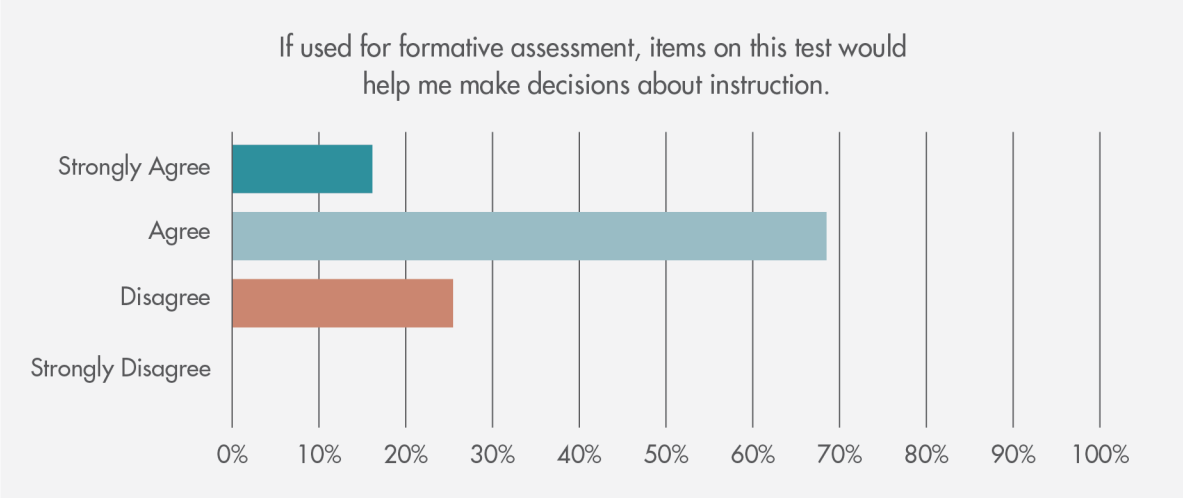


Figure 10. Percent agreement with statement: "If used for formative assessment, items on this test would help me make decisions about instruction."

In addition, panelists were generally positive about the amount of information the Smarter Balanced assessment would provide for mid-performing students. This may be attributed to the fact that the test form examined was at the 60th percentile, which is slightly above the mid-point of student performance. (Response data are presented in Figures 11 and 12.) These data are most relevant to key question 4, regarding the information provided for a range of performers.

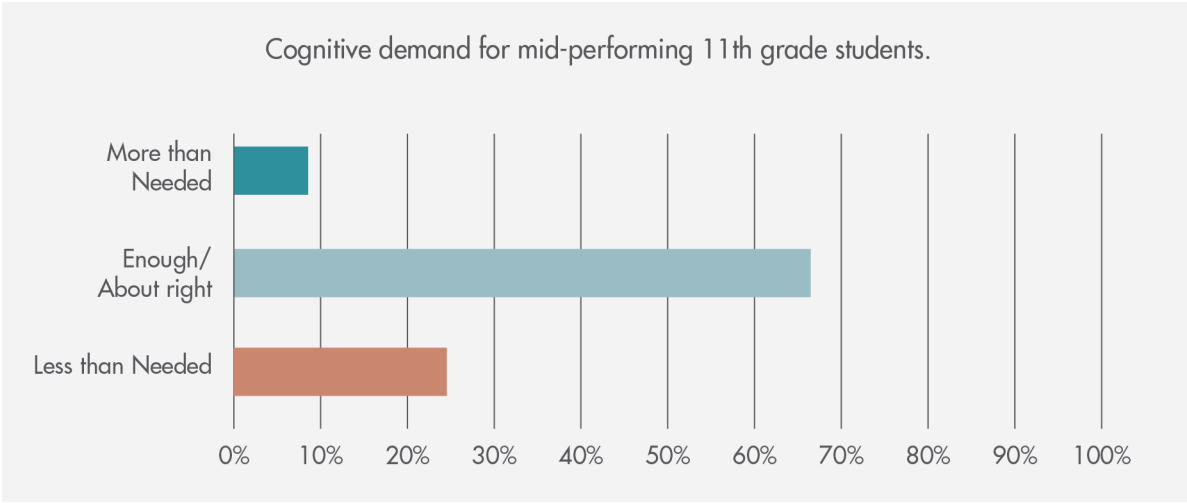


Figure 11. Percent of teachers who indicated the number of 11th grade items with "Cognitive demand for mid-performing 11th grade students" is "more than needed"; "about right/enough" or "less than needed."

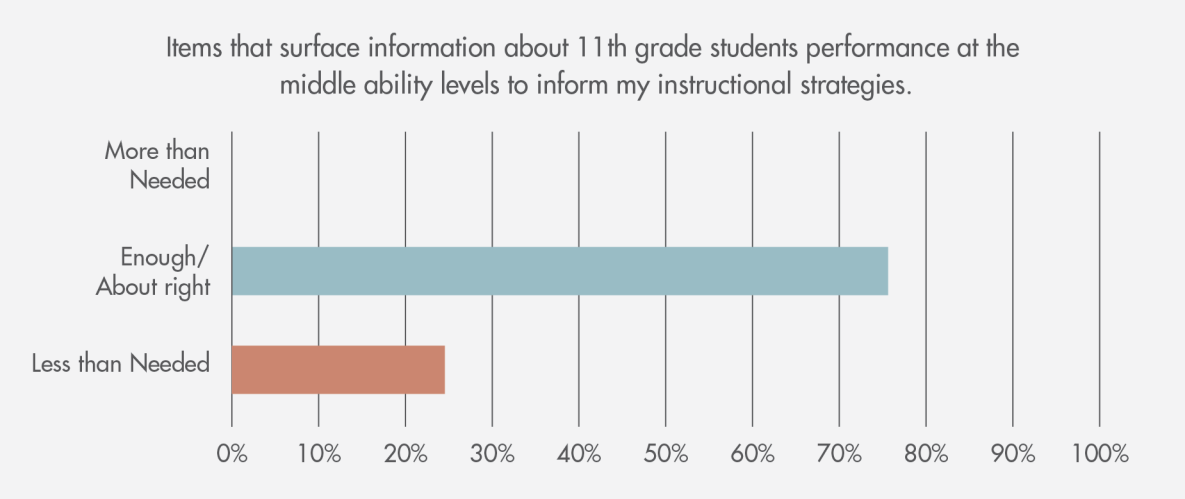


Figure 12. Percent of teachers who said that the "Items that surface information about 11th grade student performance at the middle ability levels to inform my instructional strategies" is "more than needed"; "about right/enough" or "less than needed."

Mixed and Neutral Opinions

Items in this section garnered results that were closely split between agreement and disagreement, or they received responses in every category with some frequency. For example, the panelists had split opinions—with about half agreeing and half disagreeing in both ELA and math—about the extent to which students were required to integrate information, either from within a domain of content or across domains. These results are shown in Figures 13 and 14.

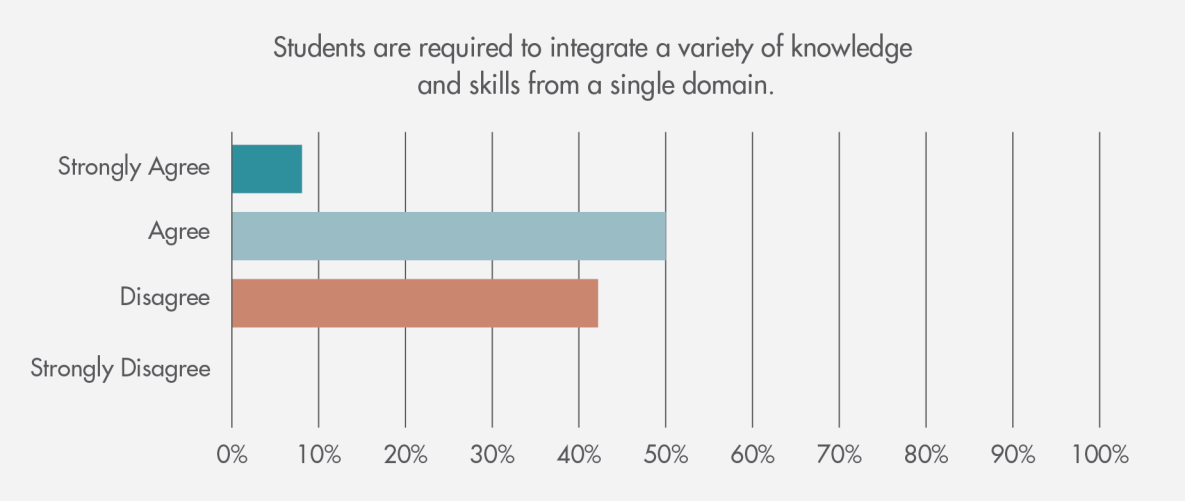


Figure 13. Percent agreement with statement: "Students are required to integrate a variety of knowledge and skills from a single domain."

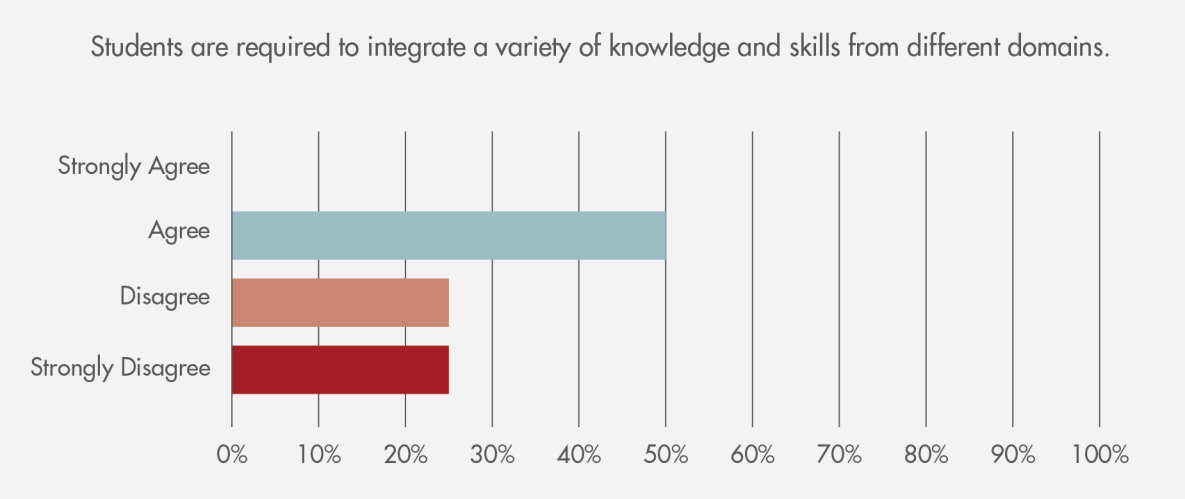


Figure 14. Percent agreement with statement: "Students are required to integrate a variety of knowledge and skills from different domains."

One characteristic that provoked mixed responses had to do with the number of items above 11th grade level. While half the teachers thought that the number of items above grade level is about right, the other half were split between there being too many or not enough. These responses are most closely aligned to key question 5.

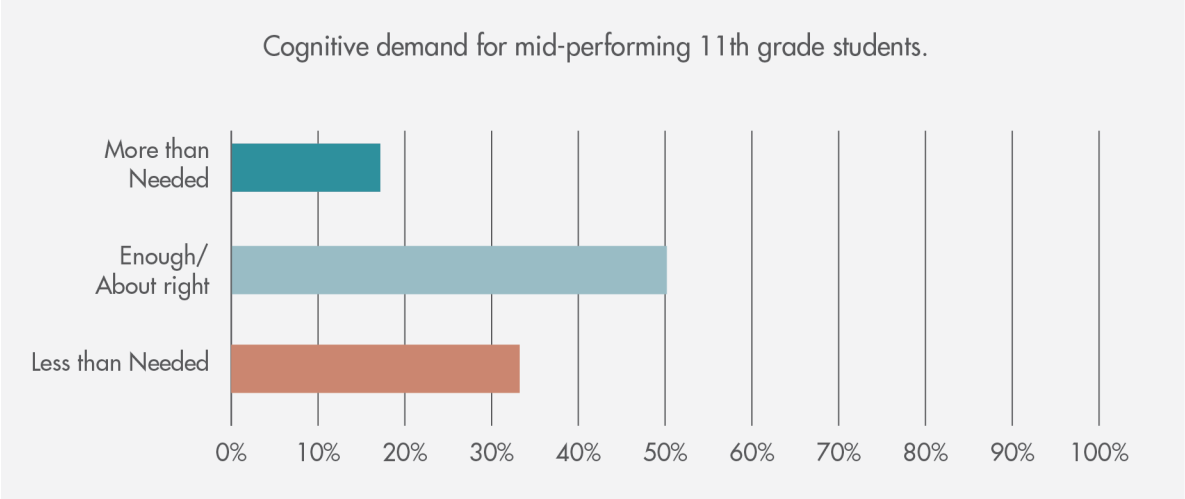


Figure 15. Percent of teachers who indicated "The number of items that are above 11th grade level" was "more than needed"; "about right/enough" or "less than needed."

Teachers also were very evenly split on whether or not the Smarter Balanced assessment emphasized certain item types more heavily than others, as shown in Figure 16. "Item types" here refers to the kinds of questions that are asked, such as multiple choice or constructed response.

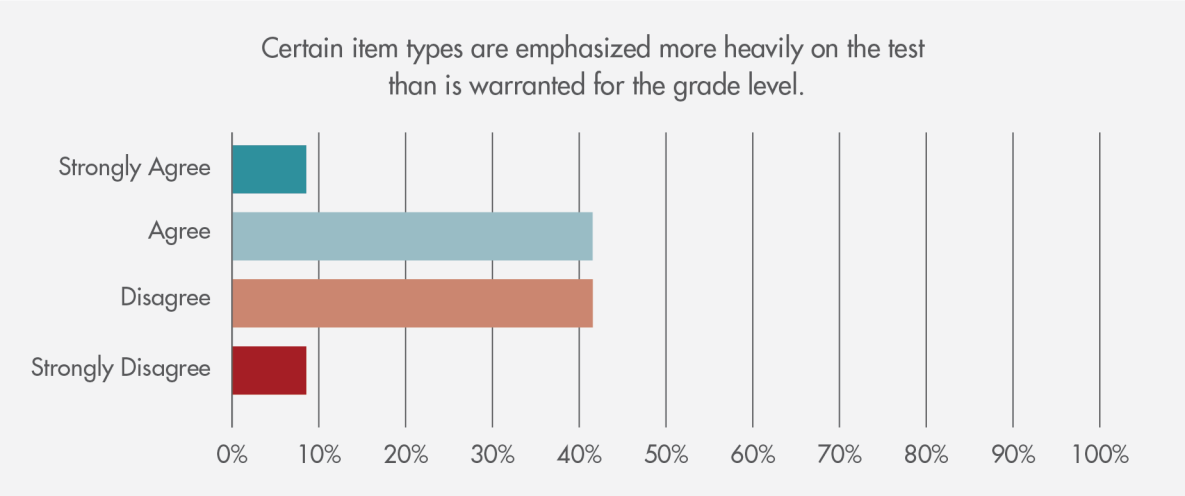


Figure 16. Percent agreement with statement: "Certain item types are emphasized more heavily on the test than is warranted for the grade level."

Critical Reactions (Areas to Improve)

The panelists stated that one area where the Smarter Balanced assessment could be improved is in the level of challenge and cognitive complexity. These data are related to key questions 2 and 4. For example, the teacher panel narrowly found that the cognitive challenge for high performing students was insufficient on the test form they reviewed, as shown in Figure 17. The teachers were split on this item, with a substantial minority believing the demand level was about right. They believed more consistently that students were not required to demonstrate complex thinking by the items on the assessment, as shown in Figure 18.

Recall that the Smarter Balanced assessment is adaptive, and the form reviewed was fixed at the 60th percentile. Student at a higher percentile rank would respond to different test items that may be more challenging.

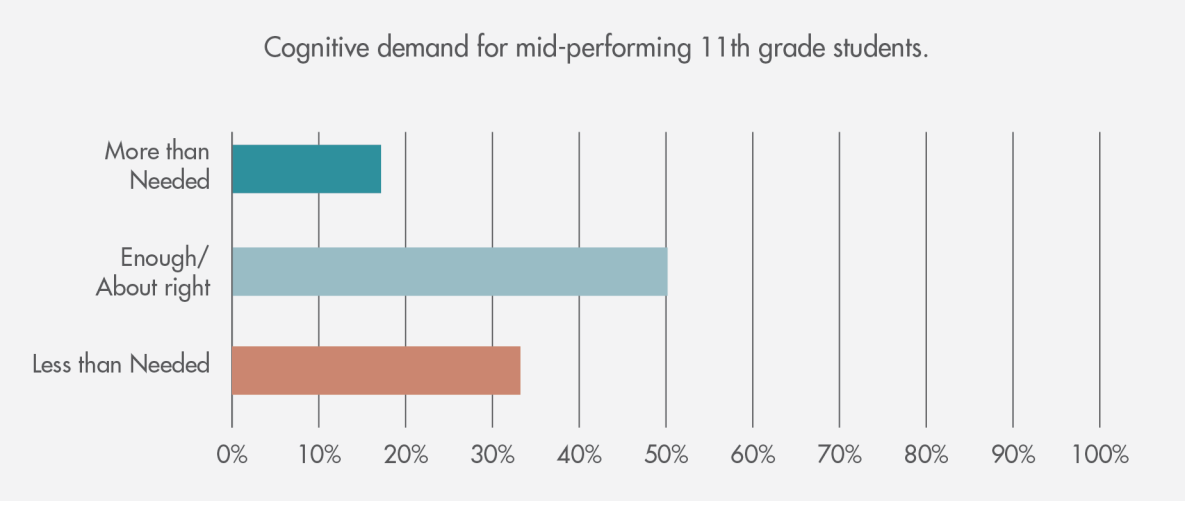


Figure 17. Percent of teachers who indicated "Cognitive demand for high-performing 11th grade students" is "more than needed"; "about right/enough" or "less than needed."

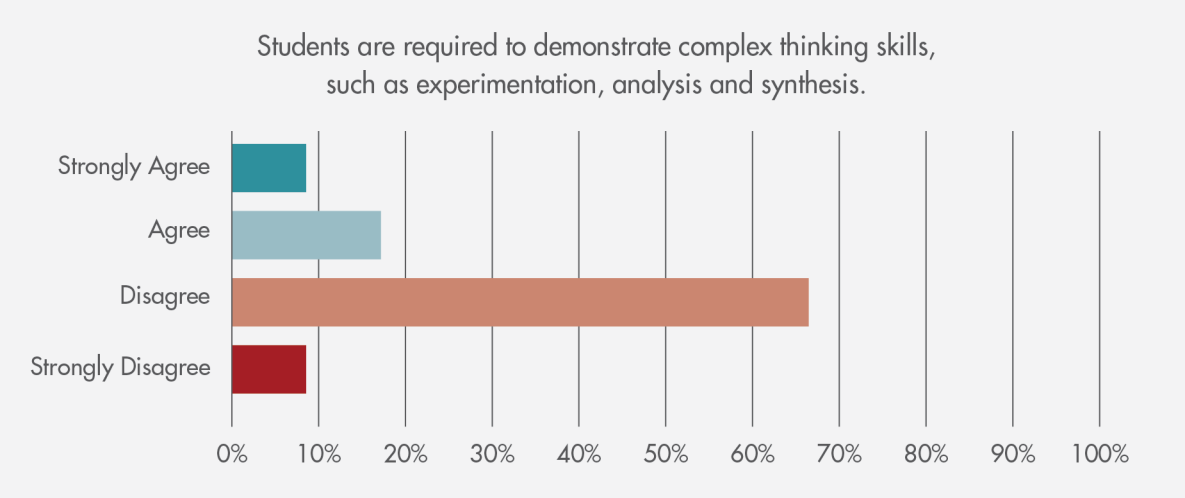


Figure 18. Percent agreement with statement: "Students are required to demonstrate complex thinking skills, such as experimentation, analysis and synthesis."

Teachers also agreed that the test would be easy for high-performing students.

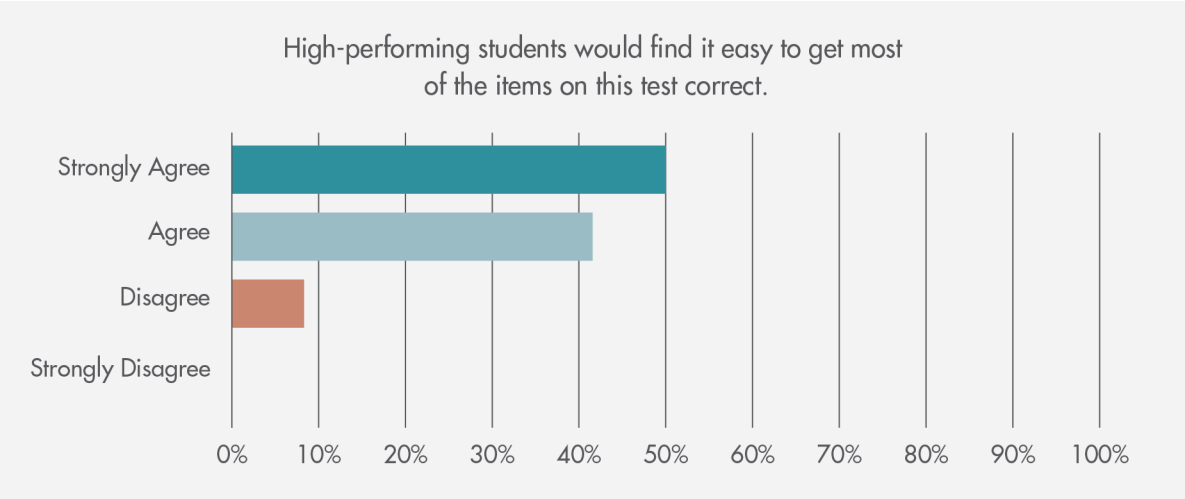


Figure 19. Percent agreement with statement: "High-performing students would find it easy to get most of the items on this test correct."

Despite relatively positive responses to questions about the depth and range of content, the panelists had some concerns that not enough emphasis is given to certain areas on the assessment, as shown in Figures 20 and 21. These items display data related to key question 3 about alignment with panelist's instructional practices.

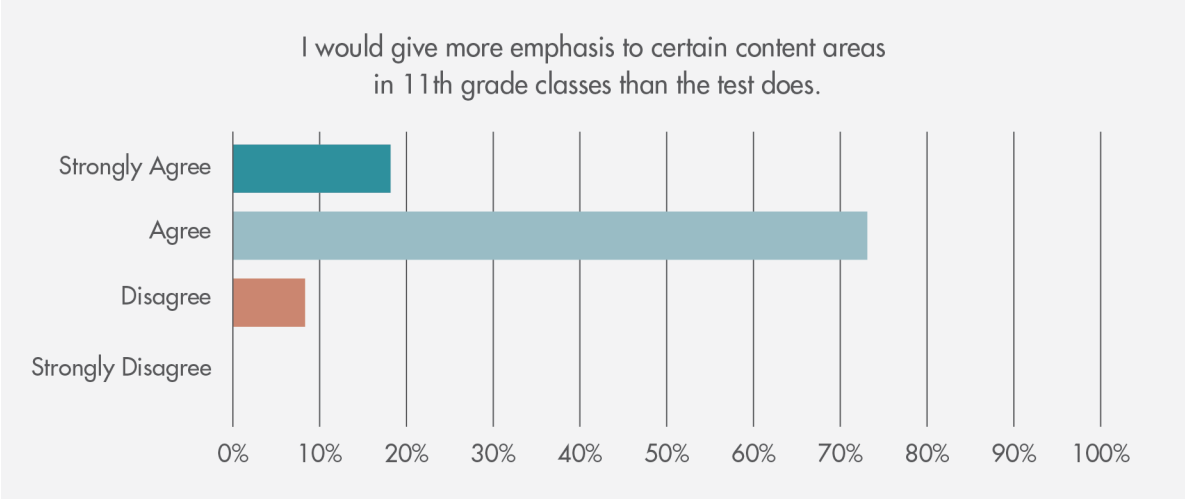


Figure 20. Percent agreement with statement: "I would give more emphasis to certain content areas in 11th grade classes than the test does."

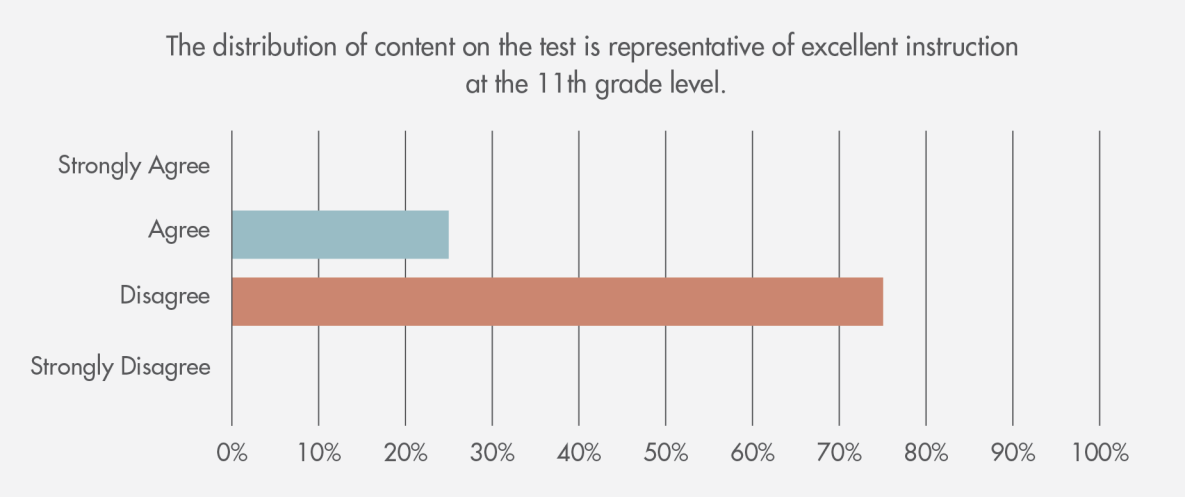


Figure 21. Percent agreement with statement: "The distribution of content on the test is representative of excellent instruction at the 11th grade level."

Even with positive ratings for other items on formative assessment, teachers did not feel that the concept coverage on the Smarter Balanced test was ideal, with 75% indicating that they believe there are gaps.

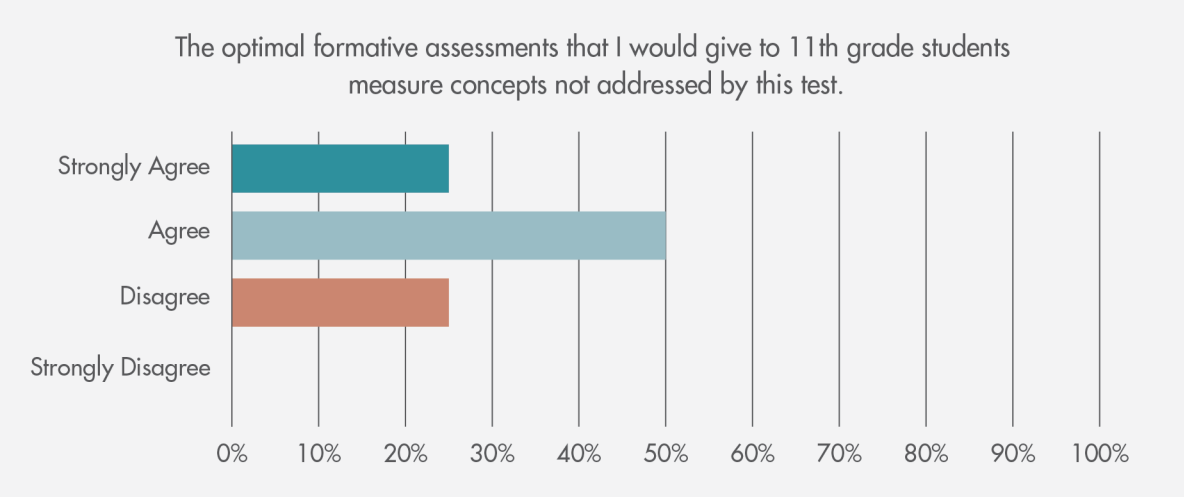


Figure 22. Percent agreement with statement: "The optimal formative assessments that I would give to 11th grade students measure concepts not addressed by this test."

Teachers were asked to rate their agreement with the statement, "This test measures the most important knowledge and skills to be taught in an excellent 11th grade math/ELA classroom." As shown in Figure 23, only 25% of the teachers on that panel strongly agreed or agreed that the 11th grade Smarter Balanced assessment measures the most important knowledge and skills taught in an excellent 11th grade math/ELA classroom. This is most relevant to key question 3. Teachers' comments suggest the assessment does not reach far enough into the higher cognitive levels to adequately measure 11th grade instruction. One teacher said, "Excellent instruction goes way deeper than this."

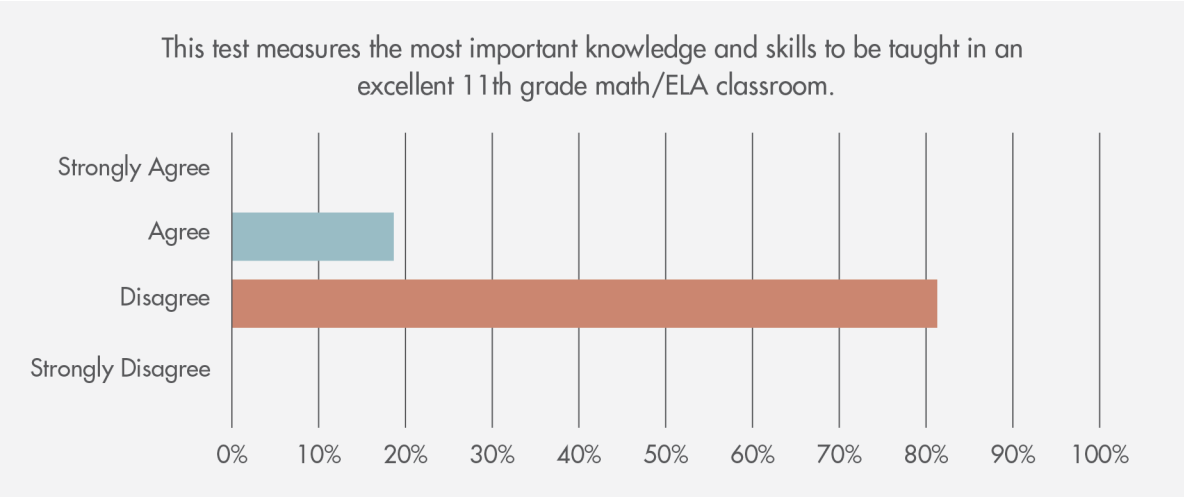


Figure 23. Percent agreement with statement: "This test measures the most important knowledge and skills to be taught in an excellent 11th grade math/ELA classroom."

Recommendations for Continuous Improvement

The findings from our study include a mixture of strengths and suggestions for continuous improvement of the 11th grade Smarter Balanced assessment. A major theme that emerged from the data for math and ELA was that the 11th grade consortium assessment generally does not reflect the kinds of content and instruction teachers would expect in excellent 11th grade classrooms. One math teacher noted, "I think that there are a lot of things missing." The panelists believed that the Smarter Balanced assessment could and should demand more of 11th grade students in terms of cognitive complexity, reasoning and communication.

Teachers' general perceptions were that more needed to be done in the classroom, in addition to content-based instruction, to prepare 11th grade students to be successful on a test like Smarter Balanced. They had suggestions for Smarter Balanced to consider as the test evolves.

More than one panelist stated that there is an over-emphasis in the ELA test on locating the main idea in a text. They believe that in the 11th grade more focus should be given to analysis and synthesis of ideas.

> I felt like there were too many questions with the nonfiction where it would ask questions like, "Pick out the main idea and the supporting evidence." On the literature we focus a lot more--not necessarily on information that's explicit--but more implicit. And I felt like a lot with the scientific [passages] … the answer you're generating [is] the main idea, but it's really just right there at the text.

> I felt like with Common Core … it affects more communication and collaboration and synthesizing multiple text[s] and coming up with original ideas and support. And with our nonfiction text, we're not looking to find just that explicit main idea—or I don't teach that way.

Another panelist noted that the structure of the test items in this specific item type might make the questions more challenging for students than intended, given the skill being assessed. That is, given the skill being assessed, the test items appear to approach the skill differently than typically required in a classroom assignment.

> The types of logic that [has] to go through the kid's mind to actually reason through and identify and select the best support for a claim, that's far different than teaching a student to write an argument with claim and support in a linear fashion. They're almost having to work backwards and around the side and I think that that logical demand is far above what that 60th percentile 11th grader can typically do.

The math teachers were positive about the requirement that students create responses such as graphs and plots as part of the assessment. This approach seems to align with skills required in their 11th grade classrooms.

> So it's not just the typical "here's a histogram, pick the right one." Or "here's a dot plot, pick the right one." But they're actually engaging in creating it and I loved the piecewise function with the videos. I thought that was awesome.

The panelists also said there is insufficient emphasis on mathematical reasoning and explication.

> One thing that I worry a little bit about is I wanted to see better critiquing and the reasoning of others on the math side. So a lot of problems were, okay, Johnnie thinks this and Janie thinks this. Which one's right and why? Well, we want to go beyond just snagging for the one mistake or the wrong answer. We want them to

> actually delve in and maybe say, well, if Johnnie's correct and Janie's not, what can Johnny say to her to steer her down the right path?

One recurring theme in the discussion concerns the information that is returned to the teachers, students, parents and schools. While the panel did not have sample data analyses or reports from the Smarter Balanced assessment to review, this topic was one of strong concern. One teacher indicated:

> I think this kind of testing is a great tool in terms of information to teachers if you can get some very specific information back to teachers and information to students and parents… specific about what an individual [student] actually knows and can do.

Another said:

> I have yet to receive a report that's truly helpful in understanding where students' weaknesses and strengths lie. Is it just a matter of they erred in a simple mathematical calculation, or was it legitimately that they did not understand the concept flat out? … we very rarely receive reports that are specific enough to help us to really narrow in on is it a strength or weakness in my teaching, or is it just a lack of ability to understand the question in the way they've worded and phrased it, or computational skills?

And they noted an important point that is not specific to the Smarter Balanced assessment, but is larger, reflecting misunderstanding of the distinction between academic standards adopted by a state or jurisdiction and the associated assessments.

> …there's a lot of confusion between any state set of standards or a Common Core and assessment. And they get lumped together, which is not correct. … educators are the only people who really get that and understand that.

This suggests that the broader education community needs to do a better job clarifying this difference for stakeholders who may not be as intimately involved as teachers in these efforts.

Concluding Thoughts

Through the insight and expertise of excellent teachers, we sought data and evidence, using five key questions to evaluate three claims we wanted excellent teachers to support or refute about the new state assessments:

1. The new consortium test supports excellent 11th grade instruction.
2. The new consortium test reflects great teaching.
3. The new consortium test is of high quality and worth the transition.

We evaluated the 11th grade Smarter Balanced assessment. While the design did not allow for comparison with other state assessments, we were still able to make inferences about its quality based on our teachers' evaluation. The assessment represents grade-level appropriate content, in both range and depth. Success on the assessment would require intentional instruction in the content areas assessed, and not "skill-and-drill" practice.

However, the teachers found that assessment content needs to be elevated to the level of rigor and challenge our teachers think is typical of excellent 11th grade instruction. Without the benefit of reviewing other state assessments, it is less clear whether the 11th grade Smarter Balanced is on the right trajectory. However, with continued development, careful implementation, strong support and training for teachers, transparency and effective communication, and patience from all stakeholder communities, the transition to the 11th grade consortium assessment should be worthwhile.

Appendix A: Panel Demographics

In this appendix, the details of the panel demographics are provided.

Figure A1: Gender

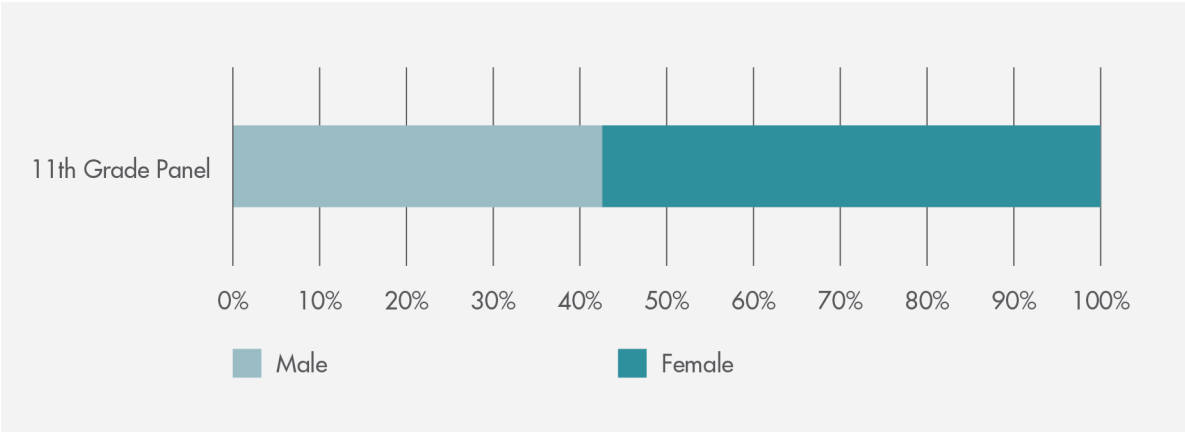


Figure A2: Race/Ethnicity

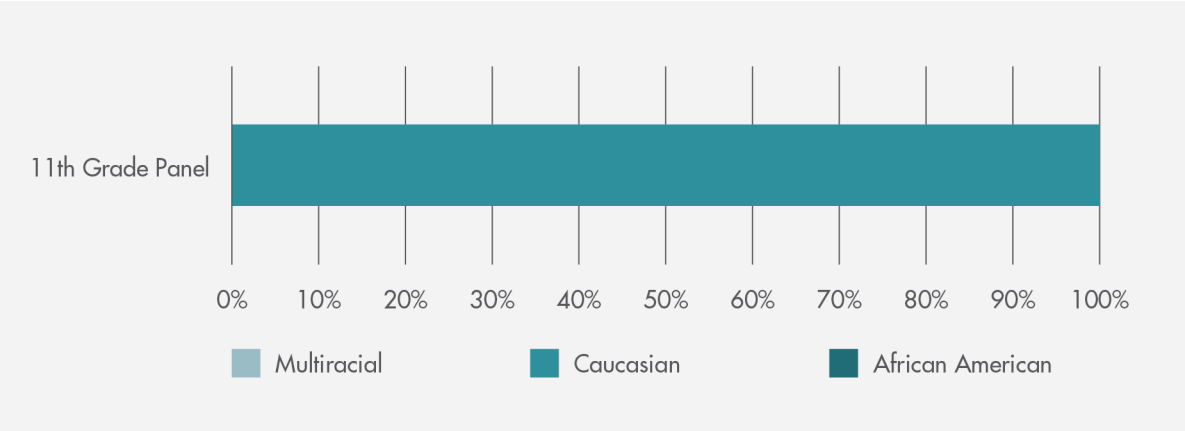


Figure A3: Years of Teaching Experience

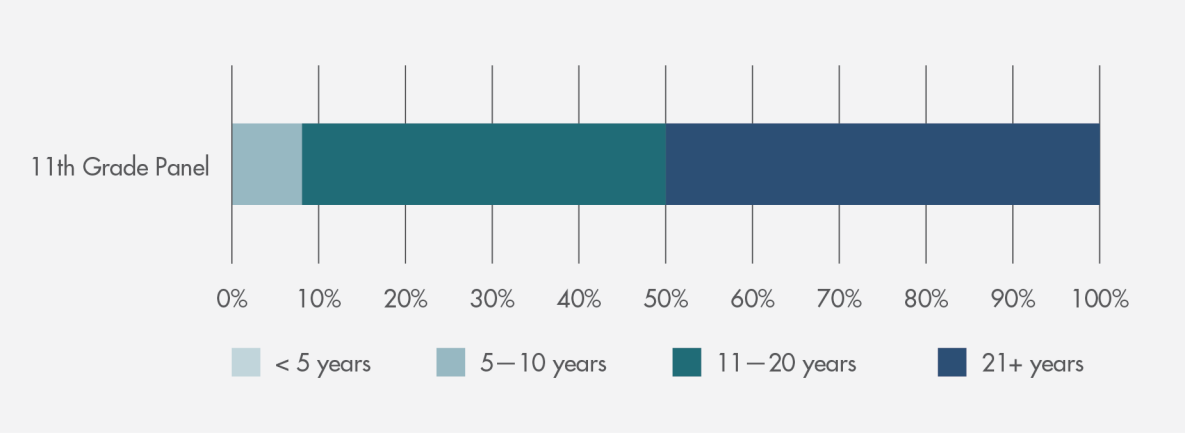
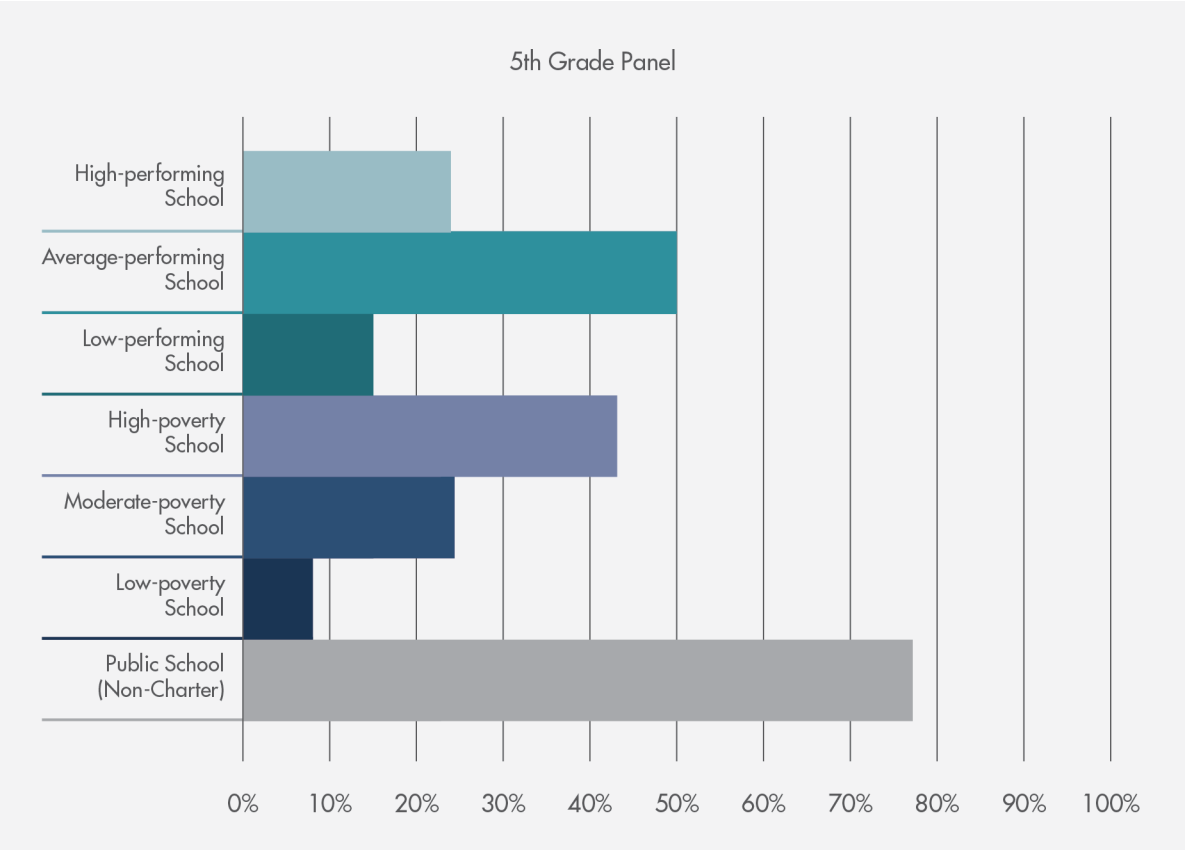


Figure A4. Teaching contexts



Appendix B:
Survey of Assessment Quality Items

Participants were asked to evaluate whether, in their judgment as an expert teacher, the assessments had "enough" of the quantity being described in the survey item below. The response scale was: "More than needed"; "Enough/About right" and "Less than needed." The results are presented below in Table B1 in two formats. The percentage of teachers who responded in each category for each assessment is shown. The percentages are shaded so that values of 50% or greater are blue.

In addition, the categories were coded as follows:

- More than needed = 3
- Enough/About right = 2
- Less than needed = 1

These values were averaged and the mean score is shown in Table B1 for the 11th grade Smarter Balanced assessment as well.

Table B1. "Amount" Items: 11th Grade Smarter Balanced

"Amount" items	SMARTER BALANCED			
	Less than Needed	Enough/ About right	More than Needed	Mean Score (1 to 3)
Items that require recall, such as identification, labeling, calculating, defining and reciting.	0%	92%	8%	2.1
Items that require application of skills, such as graphing, categorizing, organizing, predicting and estimating	8%	25%	17%	2.1
Items that require students to demonstrate strategic and extended thinking skills, such as investigation, analysis and design	58%	67%	8%	1.5
Cognitive demand for low-performing 11th grade students	17%	42%	58%	2.4
Cognitive demand for mid-performing 11th grade students	25%	67%	8%	1.8
Cognitive demand for high-performing 11th grade students	58%	42%	0%	1.4
Items that require 11th grade students to demonstrate basic knowledge of concepts	25%	58%	17%	1.9

Table B1. "Amount" Items: 11th Grade Smarter Balanced (continued)

"Amount" items	SMARTER BALANCED			
	Less than Needed	Enough/ About right	More than Needed	Mean Score (1 to 3)
Items that surface information about 11th grade student performance at the lower ability levels to inform my instructional strategies	27%	55%	18%	1.9
Items that low-performing 11th grade students would be expected to get right	50%	50%	0%	1.5
Items that low-performing 11th grade students would be expected to get wrong	8%	50%	42%	2.3
Items that surface information about 11th grade student performance at the middle ability levels to inform my instructional strategies	25%	75%	0%	1.8
Items that mid-performing 11th grade students would be expected to get right.	18%	73%	9%	1.9
Items the mid-performing 11th grade students would be expected to get wrong	27%	55%	18%	1.9
Items that surface information about 11th grade student performance at the high ability levels to inform my instructional strategies	58%	33%	8%	1.5
Items that high-performing 11th grade students would be expected to get right	8%	50%	42%	2.3
Items that high-performing 11th grade students would be expected to get wrong	33%	67%	0%	1.7
Number of items that require application of skills needed to distinguish mid-performing from low-performing 11th grade students	17%	75%	8%	1.9
Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing 11th grade students	58%	33%	8%	1.5
The number of items that are above 11th grade-level	33%	50%	17%	1.8
The number of items that are below 11th grade-level	8%	67%	25%	2.2
Items that are likely to authentically engage student interest	42%	58%	0%	1.6

Participants were asked to evaluate whether they "agreed" with statements describing the assessments in various ways in the survey item. The response scale was: "Strongly agree"; "Agree"; "Disagree" and "Strongly disagree." The results are presented below in Table B2 in the same two formats as above and with the same shading protocol, where percentages are shaded with values of 50% or greater in blue. The categories were coded as follows:

- Strongly agree = 4
- Agree = 3
- Disagree = 2
- Strongly disagree = 1

These values were averaged and the mean score is shown in Table B2 for each assessment.

Table B2. "Agree" Items: 11th Grade Smarter Balanced

"Agree" Items	SMARTER BALANCED				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Mean Score (1 to 4)
Students are required to integrate a variety of knowledge and skills from a single domain.	0%	42%	50%	8%	2.7
Students are required to transfer knowledge from different domains.	8%	50%	42%	0%	2.3
Students are required to integrate a variety of knowledge and skills from different domains.	25%	25%	50%	0%	2.3
This test provides sufficient opportunity to evaluate students' ability to communicate in writing.	42%	33%	25%	0%	1.8
This test provides sufficient opportunity to evaluate students' ability to show their reasoning when solving a problem or arguing a case.	17%	42%	42%	0%	2.3
This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills.	8%	42%	50%	0%	2.4
Students are required to demonstrate complex thinking skills, such as experimentation, analysis and synthesis.	8%	67%	17%	8%	2.3

Table B2. "Agree" Items: 11th Grade Smarter Balanced (continued)

"Agree" Items	SMARTER BALANCED				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Mean Score (1 to 4)
This test is more cognitively demanding than is warranted for the 11th grade level.	36%	36%	9%	18%	2.1
This test is less cognitively demanding than is warranted for the 11th grade level.	33%	33%	33%	0%	2.0
Items on this test are consistent with what excellent 11th grade math/ELA teachers ask their students to know and do.	8%	50%	42%	0%	2.3
Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice.	0%	33%	33%	33%	3.0
One criterion for a high-quality assessment is that the assessment allows students to transfer their learning to new situations and problems. This test meets that criterion.	17%	33%	42%	8%	2.4
This test measures an appropriately broad sampling of the ELA/math knowledge and skills in instruction an excellent 11th grade classroom.	17%	58%	25%	0%	2.1
Excellent 11th grade instruction generally aligns with the content measured on this test.	0%	50%	42%	8%	2.6
This test measures the most important knowledge and skills to be taught in an excellent 11th grade math/ELA classroom.	0%	82%	18%	0%	2.2
This test measures the learning outcomes that I would set for student learning in 11th grade classes.	17%	42%	33%	8%	2.3
Certain item types are emphasized more heavily on the test than is warranted for the grade level.	8%	42%	42%	8%	2.5
Certain content areas are emphasized more heavily on the test than is warranted for the grade level.	9%	45%	45%	0%	2.4
I would give more emphasis to certain content areas in 11th grade classes than the test does.	0%	9%	73%	18%	3.1
The distribution of content on the test is representative of excellent instruction at the 11th grade level.	0%	75%	25%	0%	2.3

Table B2. "Agree" Items: 11th Grade Smarter Balanced (continued)

"Agree" Items	SMARTER BALANCED				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Mean Score (1 to 4)
The depth of content represented on the test is grade-level appropriate.	0%	33%	67%	0%	2.7
The range of content represented on the test is grade-level appropriate.	0%	33%	67%	0%	2.7
One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion.	8%	8%	83%	0%	2.8
If I backwards-mapped a 11th grade lesson against items like those on this test, it would help inform my lesson plan and guide me toward high quality instruction.	8%	17%	42%	33%	3.0
I would like to use formative assessments built using items from this test in a 11th grade classroom.	0%	25%	58%	17%	2.9
The optimal formative assessments that I would give to 11th grade students measure concepts not addressed by this test.	0%	25%	50%	25%	3.0
If used for formative assessment, items on this test would help me make decisions about instruction.	0%	25%	50%	25%	3.0
Student results from this test would give me valuable information about how students are learning.	9%	36%	45%	9%	2.5
The item types on this test are aligned with the skills they appear to be designed to measure.	0%	17%	67%	17%	3.0
This test provides a satisfactory balance between selected-response items and constructed response/performance-based items.	25%	17%	58%	0%	2.3
Low-performing students would find it easy to get most of the items on this test correct.	42%	50%	8%	0%	1.7

Table B2. "Agree" Items: 11th Grade Smarter Balanced (continued)

"Agree" Items	SMARTER BALANCED				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Mean Score (1 to 4)
Mid-performing students would find it easy to get most of the items on this test correct.	8%	25%	58%	8%	2.7
High-performing students would find it easy to get most of the items on this test correct.	0%	8%	42%	50%	3.4
Low-performing students would generally perform well on this test.	42%	42%	17%	0%	1.8
Mid-performing students would generally perform well on this test.	8%	17%	67%	8%	2.8
High-performing students would generally perform well on this test.	0%	8%	25%	67%	3.6
Students would likely be authentically engaged in items from this test.	17%	42%	42%	0%	2.3

Appendix C: Guiding Questions for Panel Discussions

A set of standard questions was developed based on the survey data, and follow-up prompts were incorporated organically throughout the discussion. The standard questions asked of each panel are listed below.

1. Were there any aspects of the study that may have prejudiced your judgments in favor of one test or another before you started today's survey?
2. Were there any aspects of the study that may have prejudiced your judgments in favor of one test or another while you were completing the survey?

The next set of questions were motivated by the panel's survey data:

1. A majority of you of responded that Smarter Balanced assessment contained enough items that authentically engage student interest. What are some of the ways the assessment achieves this? How can the assessment be more authentically engaging?
2. A number of you disagreed or strongly disagreed with the statement, "Items on this test are consistent with what excellent 11th grade math/ELA teachers ask their students to know and do." What are excellent 11th grade teachers asking students to do that this assessment does not capture?
    a. What knowledge and skills are missing from the assessment?
    b. What are the important knowledge and skills that are missing from the assessment?
3. What aspects of this study will you be taking away with you today?
4. What are you going to do with the information that was shared with you during this study?

Appendix D: *Attitudes Toward Test*

Teachers were given an *Attitudes Toward Tests* survey to measure shifts in their perceptions of tests and test items over the course of the study. As shown in Table 1, the largest differences (.30 of a point or greater) or change in mean scores were for the statement: "Selected-response tests are simply easier to administer than constructed-response or performance-based tests." Teachers on the 11th grade panel agreed more with this statement after evaluating the assessments

Table D1. Average *Attitudes Toward Tests* Results for 11th Grade Panel. Detail may not sum to total due to rounding.

Pre-Mean (1 to 4)	Attitudes toward Tests items	Post-Mean (1 to 4)	Pre-Post Difference
1.8	I prefer tests that are comprised mostly of selected-response items	1.8	0.0
1.8	Tests that are largely selected-response are more appropriate for the knowledge and skills embedded in my learning outcomes than constructed-response or performance-based tests.	1.7	- 0.1
3.2	I prefer tests that are comprised mostly of constructed-response or performance-based items.	3.3	0.1
3.3	Tests that are largely constructed-response/performance based are more appropriate for the knowledge and skills embedded in my learning outcomes than selected-response tests.	3.3	0.0
3.2	I prefer tests with some selected-response and some constructed-response items.	3.1	- 0.1
3.2	Tests that are comprised of some selected-response items and some constructed-response items are more appropriate for the knowledge and skills embedded in my learning outcomes than multiple-choice tests.	3.4	0.2
2.9	Selected-response tests are simply easier to administer than constructed-response or performance-based tests.	3.2	0.3
2.8	Selected-response items can be used to measure complex thinking skills.	2.8	0.1


NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY